DOCUMENT RESUME

ED 478 096 IR 058 690

| | |
|---|---|
| AUTHOR | Rocker, JoAnne; Roncaglia, George J.; Heimerl, Lynn N.; Nelson, Michael L. |
| TITLE | The NASA Scientific and Technical Information (STI) Program's Implementation of Open Archives Initiative (OAI) for Data Interoperability and Data Exchange. |
| PUB DATE | 2002-06-00 |
| NOTE | 11p.; In: SLA 2002: Putting Knowledge to Work. Papers Presented at the Special Libraries Association Conference (Los Angeles, California, June 9-12, 2002); see IR 058 674. |
| AVAILABLE FROM | For full text: http://www.sla.org/content/Events/conference/2002annual/ confpap2002/papers2002conf.cfm/ . |
| PUB TYPE | Reports - Descriptive (141) -- Speeches/Meeting Papers (150) |
| EDRS PRICE | EDRS Price MF01/PC01 Plus Postage. |
| DESCRIPTORS | Agency Cooperation; Cooperative Programs; *Data Collection; Federal Government; Government Publications ; *Information Dissemination; *Information Transfer; *Scientific and Technical Information |
| IDENTIFIERS | *Interoperability; National Aeronautics and Space Administration |

ABSTRACT

Interoperability and data-exchange are critical for the survival of government information management programs. E-government initiatives are transforming the way the government interacts with the public. More information is to be made available through Web-enabled technologies. Programs such as the NASA's Scientific and Technical Information (STI) Program Office are tasked to find more effective ways to disseminate information to the public. The NASA STI Program is an agency-wide program charged with gathering, organizing, storing, and disseminating NASA-produced information for research and public use. The program is investigating the use of a new protocol called the Open Archives Initiative (OAI) as a means to improve data interoperability and data collection. OAI promotes the use of the OAI harvesting protocol as a simple way for data sharing among repositories. In two separate initiatives, the STI Program is implementing OAI. In collaboration with the Air Force, Department of Energy, and Old Dominion University, the NASA STI Program is participating in the Technical Report Interchange (TRI) Project to exchange data between the three organizations. The second initiative is the deployment of OAI for the NASA Technical Report Server (TRS) environment. The NASA TRS environment is comprised of distributed technical report servers with a centralized search interface. This paper focuses on the implementation of OAI to promote interoperability among diverse data repositories. (Contains 10 references and 4 figures.) (Author/MES)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 478 096

# The NASA Scientific and Technical Information (STI) Program's Implementation of Open Archives Initiative (OAI) for Data Interoperability and Data Exchange*

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

E. Williams

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

JoAnne Rocker
George J. Roncaglia
Lynn N. Heimerl
Michael L. Nelson
NASA Langley Research Center, Hampton, VA

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

## ABSTRACT

Interoperability and data-exchange are critical for the survival of government information management programs. E-government initiatives are transforming the way the government interacts with the public. More information is to be made available through web-enabled technologies. Programs such as the NASA's Scientific and Technical Information (STI) Program Office are tasked to find more effective ways to disseminate information to the public. The NASA STI Program is an agency-wide program charged with gathering, organizing, storing, and disseminating NASA-produced information for research and public use. The program is investigating the use of a new protocol called the Open Archives Initiative (OAI) as a means to improve data interoperability and data collection. OAI promotes the use of the OAI harvesting protocol as a simple way for data sharing among repositories. In two separate initiatives, the STI Program is implementing OAI. In collaboration with the Air Force, Department of Energy, and Old Dominion University, the NASA STI Program has funded research on implementing the OAI to exchange data between the three organizations. The second initiative is the deployment of OAI for the NASA technical report server (TRS) environment. The NASA TRS environment is comprised of distributed technical report servers with a centralized search interface. This paper focuses on the implementation of OAI to promote interoperatibility among diverse data repositories.

## NASA STI PROGRAM OFFICE

The Scientific and Technical Information (STI) Program Office has existed since the early days of the National Aeronautics and Space Administration. Its purpose, by statute, is inherent in

IR058690

* The work entitled , "The NASA Scientific and Technical Information (STI) Program's Implementation of Open Archives Initiative (OAI) for Data Interoperability and Data Exchange," was prepared as part of my official duties as an employee of the U.S. Government and, in accordance with 17 U.S.C. 105, is not available for copyright protection in the United States.



BEST COPY AVAILABLE

NASA's mission, as defined in the National Aeronautics and Space Act of 1958, is "...to provide for the widest practicable and appropriate dissemination of information concerning its activities and the results thereof. "

The STI Program ensures that NASA remains at the leading edge of R&D by quickly and efficiently capturing worldwide scientific and technical information for use in problem solving, awareness, and knowledge transfer. Its data collection and dissemination supports NASA's mission to communicate scientific knowledge and understanding by collecting and transferring NASA's research and development (R&D) to the aerospace and academic communities. The program collects the NASA-produced information from 10 NASA Centers and Headquarters, other sources in the U.S., and over 50 foreign countries, and maintains access to the largest collection of aerospace science and technical information in the world.

Some of the STI Program activities include the following, which are to:

- Collect, announce, disseminate, and archive all STI resulting from NASA-funded and sponsored research to reduce duplication of effort and improve productivity and cost-effectiveness of the NASA research effort
- Acquire domestic and international STI pertinent to NASA's missions
- Handle and publish all appropriately reviewed STI for NASA, thus requiring close coordination with export control, patent, copyright, and intellectual property organizations, and international partnerships
- Build and maintain the STI Database
- Coordinate the Agency's various Field Center STI programs
- Develop and implement all Agency policy and procedures for external release of STI
- Monitor a contractor facility, the NASA Center for AeroSpace Information, in Hanover, MD
- Negotiate and handle all STI international agreements (with Code I) for NASA

Over the years, the STI Program Office has faced challenges in its task to acquire and promote information usage. Advances in information technology have fueled higher expectations and increased demand for easy and efficient access to scientific information. Internet savvy users are used to sophisticated search engines that retrieve documents, photos, music, and other kinds of data. Internet users expect more from information providers:

- Desktop access to full-text documents *vs.* abstracts
- Rapid access to documents to meet customer requirements
- Wider access to varied information formats
- Preprints and other forms of gray literature not published in traditional forms (e.g., photos, videos, and graphics)
- High-speed Internet access and web-based architecture
- Better data organization and more-user friendly interfaces

New challenges to information delivery are not the only concerns of the STI Program Office. The high cost of supplementing the NASA-produced information with commercial data in order to broaden the access to scientific and technical research has limited the kinds of

information that the program can make available. Acquisition of commercial data is expensive and restrictive licensing hinders efforts by the STI Program Office to enhance its aerospace knowledge base. As the STI Program continues to face funding challenges, finding alternatives to commercially available information is necessary to ensure that NASA users have the information they need. The emergence of the Open Archives Initiative (OAI) as a technology bridge connecting heterogeneous data sources offers the STI Program a way to build its collection circumventing the problems associated with commercial data. Further, as more information providers start using OAI for data exchange, the breadth and scope of information available to the STI Program will grow.

## Open Archives Initiative

One of the main barriers to information exchange is the multitude of metadata formats used by database and archive creators. The Open Archives Initiative grew out of the belief that simplifying data exchange would increase scholarly communication (Van de Sompel and Lagoze, 2001). Costly and hard-to-implement protocols like Z39.50 were barriers to data sharing among repositories. OAI was established as a low-cost, low-barrier protocol for transferring data between archives.

OAI is a relatively new effort. The first meeting of the Open Archives Initiative was held in October 1999, in Santa Fe, NM. The Santa Fe Convention, as the meeting became known, brought together information and computer science specialists to discuss and solve issues of interoperability among electronic preprint (e-print) archives (Van de Sompel and Lagoze, 2000; Van de Sompel et al., 2000). Discussions focused on a new approach to metadata harvesting data between repositories. In this new data harvesting approach, a distinction is made between data providers and service providers. As defined by the Santa Fe Convention:

- A *data provider* is the manager of an e-print archive, acting on behalf of the authors submitting documents to the archive
- A *service provider* is a third party, creating end-user services based on data stored in e-print archives

The convention further defined an archive as a collection of records. Records have the following properties:

- A record in an e-print archive contains, at least, metadata that describes full content
- A record in an e-print archive may also contain full content such as a research paper, a dataset, software, etc. or a bundle of these (OAI, "The Santa Fe Convention")

The Santa Fe Convention did the preliminary work in developing a metadata harvesting protocol by establishing guidelines for identifying archives as data and/or service providers, specifying metadata formats, and harvesting data from data providers. These guidelines were incorporated into the first version of, "The Open Archives Initiative Protocol for Metadata Harvesting," developed in January 2001. The protocol was updated to version 1.1 in July 2001 and release of version 2 of the protocol is scheduled for May 2002.

The OAI protocol differs from other data exchange protocols such as Z39.50 (Lynch, 1997) in that it is designed for simplicity. Harvesting is initiated by HTTP-encoded queries to OAI-compliant archives and metadata is returned in XML. The protocol calls for a standard metadata format based upon the fifteen elements of Dublin Core (DCMI, 1999). The use of Dublin Core removes the burden of trying to map between multiple metadata formats. An OAI-layer can be put over existing information systems using Perl CGI scripts, Java servlets, PHP scripts, or any number of possible implementations.

OAI is an ongoing collaborative effort lead by Herbert Van de Sompel of Los Alamos National Laboratory and Carl Lagoze of Cornell University. An OAI Steering Committee sets policy and a Technical Committee continues development of the harvesting protocol. Funding and support for OAI come from the Digital Library Federation, the Coalition for Networked Information, and the National Science Foundation.

OAI is intentionally designed to be a low-cost, low-barrier approach to information interoperability. The STI Program decided to invest in this promising technology by funding two initiatives to test OAI's applicability in the NASA STI environment. If successful, these OAI projects will significantly increase accessibility of aerospace information for NASA.

## TECHNICAL REPORT INTERCHANGE (TRI) PROJECT

The federal government is one of the greatest publishers of scientific and technical information. A great portion of this information is unlimited and unclassified without any formal restrictions for its use. As costs for commercial sources of scientific and technical literature escalate, the STI Program has looked to other federal agencies with similar research programs as partners for collaborative sharing efforts.

The STI Program Office is participating in the TRI Project, a collaborative data sharing experiment with the Air Force, and Department of Energy (DOE) sites. NASA, the Air Force, DOE sites have overlapping research disciplines; therefore, individual collections of documents were of mutual interest to all parties involved. The TRI Project's primary objective is for the exchange of metadata using the OAI harvesting protocol. The project is made more complex because organizations embed links within their metadata records to the full-text image residing on institutional servers. No actual image files are exchanged between participants. Using the OAI protocol in this way means participants will be able collect metadata from each other and access full-text documents. The technical work for the TRI Project is carried out by the ODU Team, a group represented by the faculty and staff from the Computer Science Department at Old Dominion University, Norfolk, VA (URL: http://dlib.cs.odu.edu). The ODU Team created and continues to modify software code that provides the mechanism for data exchange.

### PROJECT PHASES

The TRI Project has several different phases. During Phase 1, the ODU Team developed a TRI Software Package for each participant. The project is currently in Phase 2, in which the TRI Package is installed at each participant's site. Software testing and security modifications

continue to occur in Phase 2. The final stage of the project, Phase 3, will be achieved when all the sites are fully operational. The projected completion date for the project is September 2002.

**TRI Software Development (Phase 1)**

The initial phase of the project involved developing software tools would allow the participants to exchange data. The Old Dominion University (ODU) team created a "TRI Package" for each of the sites. The package consisted of several different software components for the OAI implementation:

- An OAI-compliant repository
- Code to convert organizational metadata from its native format to Dublin Core
- Code to convert Dublin Core metadata back into native metadata format
- Harvesting code
- Harvesting scheduler code
- Log files to track when a site harvests or has been harvested.

*Figure 1* shows the TRI Project architecture for one participant. This architecture is duplicated for all the participants in the TRI Project. Each site has its library or information system repository (1). This is the system the users search for institutional data. Metadata from the repository is converted from its native format into Dublin Core (2) using conversion code written by the ODU Team. The Dublin Core data is moved into an OAI-compliant repository (3). The OAI-compliant repository is the harvesting point for outgoing (4) and incoming (5) harvest commands. Newly harvested data is converted from Dublin Core back into the institutional metadata format and ingested into the site's information system (6). Users can now search on metadata from other institutions. It is transparent to users that they are searching data from an external site because the new data in input into the system they are used to searching for information.

**TRI Project Architecture (One Participant)**

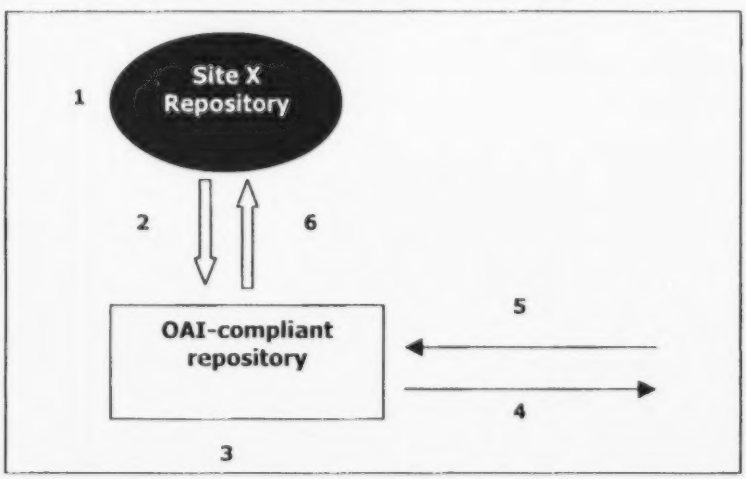


**Figure 1**

## Metadata Formats

Each organization has its own unique metadata format. Old Dominion University worked with each individual organization to develop code to map its native metadata into Dublin Core. NASA uses a metadata format that is a combination of COSATI and MARC format. *Figure 2* illustrates how a NASA converted into Dublin Core from its native format. *Figure 3* shows the same record but coded in XML.

### NASA Dublin Core Metadata Record

| | |
|---|---|
| *Source DL:* | NASA |
| *Identifier:* | http://naca.larc.nasa.gov/reports/1946/naca-rb-l5k02/ |
| *Title:* | Aerodynamic Characteristics of Four NACA Airfoil Sections Designed for Helicopter Rotor Blades |
| *Creator:* | Louis S. Stivers |
| *Creator:* | Fred J. Rice, Jr. |
| *Description:* | Four NACA airfoil sections, the NACA 7-H-12, 8-H-12, 9-H-12, and 10-H-12, suitable for use as rotor-blade sections for helicopters and other rotary-wing aircraft have been derived and tested. These airfoil sections have comparatively low drags in the range of low and moderate lifes and small pitching moments that are nearly constant up to maximum lift. |
| *Contributor:* | Langley Memorial Aeronautical Laboratory |
| *Discovery:* | February 1946 |
| *Type:* | NACA Restricted Bulletin L5K02; Wartime Report WR-L-29 |
| *ID:* | oai:NACA:1946:naca-rb-l5k02 |
| *Set:* | NACA |
| *DateStamp:* | 2001-07-27 |

**Figure 2**

### NASA Metadata Record Coded in XML (partial record)

```
<?xml version="1.0" encoding="UTF-8"?>
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/1.1/OAI_GetRecord
http://www.openarchives.org/OAI/1.1/OAI_GetRecord.xsd">
<responseDate>2002-03-11T22:01:04+00:00</responseDate>
<requestURL>http%3A%2F%2Fnaca.larc.nasa.gov%2Foai%2Findex.cgi%
3Fverb%3DGetRecord%26identifier%3Doai%3ANACA%3A1946%3Anaca-
rb-l5k02</requestURL>
<record>
<header>
<identifier>oai:NACA:1946:naca-rb-l5k02</identifier>
<datestamp>2001-07-27</datestamp>
</header>
<metadata>
<dc xmlns="http://purl.org/dc/elements/1.1/"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:schemaLocation="http://purl.org/dc/elements/1.1/
        http://www.openarchives.org/OAI/1.1/dc.xsd">
```

**Figure 3**

### TRI Deployment (Phase 2)

With the TRI Package software developed, the TRI Project moved into its current stage, Phase 2. The ODU Team is helping install the TRI Package at each site. Upon completion of installation, each institution will populate its OAI-compliant repositories with metadata to test the harvesting tools. Testing the TRI Package tools, identifying technical issues, and resolving software concerns will occur in Phase 2. Each institution will need to ensure that security measures are in place for harvesting metadata and accessing full-text documents.

### Integration and Full Operability (Phase 3)

Phase 3 is the final phase of the experiment. Each participant will fully deploy the TRI Package. New metadata content will be added to feed the repositories. The ODU Team will develop administration, deletion, and modification functionality into the TRI tools so participants can manage their metadata content. Participants will decide if they would like to continue development of the TRI Project beyond its first several software iterations. Evaluation of the project and assessment of future goals will help determine future possibilities for the TRI Project. The participants may continue with a collaborative approach to project development, or they may decide to do individual implementations of OAI.

## NASA TECHNICAL REPORT SERVERS OAI PROJECT

The second initiative sponsored by the STI Program Office is the implementation of OAI for the technical report servers (TRS) at NASA. NASA centers and STI Program Office disseminate NASA-produced information to the public through a network of technical report servers distributed across the Agency. These technical report servers provide access to bibliographic citations and, in some cases, electronic full-text documents. There are two ways users can search the TRS. They can search each TRS individually, or they can search all the TRS via the NASA Technical Report Server (NTRS) (Nelson et al., 1995). NTRS is a meta search engine that conducts search queries across all TRS nodes and displays a canonical list of results. The NTRS receives an average of 100K hits a month.

The present TRS environment supports NASA's mission to communicate its scientific research, however, the security vulnerabilities of WAIS has made it necessary to look at alternative ways of providing access to NASA's scientific and technical information. OAI's data and service provider design architecture offers a new model for data sharing and searching. Centers will become OAI-compliant without having to change their internal publishing processes and significant costs. Some scripting will be necessary to export existing metadata into Dublin Core and XML; however, once complete, OAI will not impact regular operations of the TRS.

In the proposed plans to add OAI capabilities to the TRS environment, the NTRS meta search engine will be modified to become an OAI service and data provider. This moves NTRS from being a distributed searching interface to a metadata harvesting interface. The NTRS will harvest from the TRS at each center and build its own repository of data. Centralizing of data overcomes two of the problems with the current configuration. When users search NTRS now, they do not retrieve an integrated list of results. Results list by each node so users have to scroll

through the list from each center. An integrated repository of data will allow more functionality in sorting and displaying search results. Additionally, an integrated approach to centralizing data means that NTRS performance will not be impacted by TRS downtime. When a TRS node goes down, the meta search engine gets tied up and affects system performance. A centralized repository of data eliminates the scalability problem of a distributed search across multiple nodes.

*Figure 4* shows how the TRS and NTRS will interact with each other when OAI is implemented. Users will be able to search the individual nodes as they do now. There will be an OAI-layer over the existing systems that will allow for harvesting by the NTRS server. NTRS will function as both a service provider and a data provider. Other external OAI harvesters will be able to query NTRS for its data. NTRS, in turn, will harvest from external OAI-compliant repositories and increase the amount of subject content available for internal NASA users and the public. As scientific and technical research becomes more cross disciplinary, OAI will provide a bridge between multiple research areas.

**TRS and NTRS Future Environment**

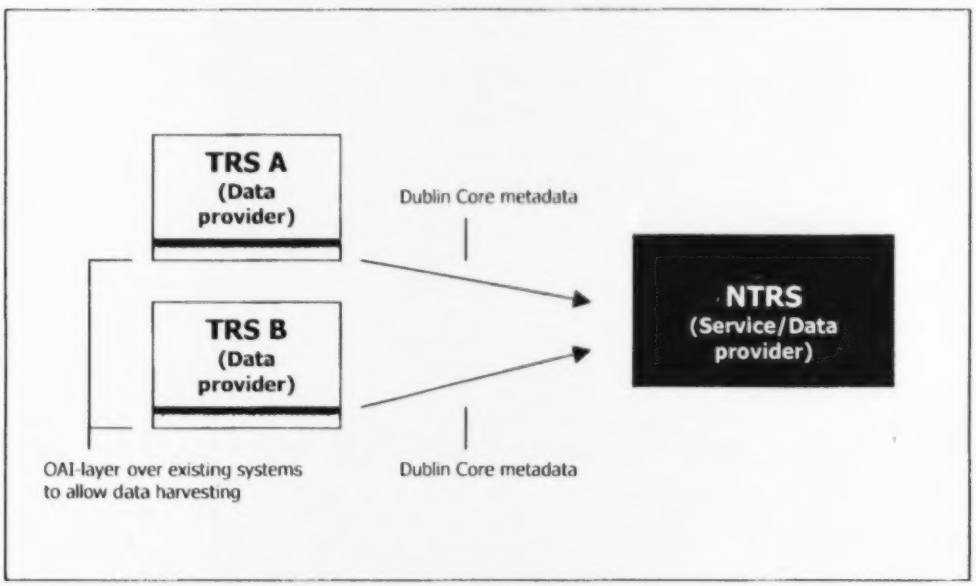


**Figure 4**

## CONCLUSION

The STI Program Office faces new challenges as the Agency enters the 21st century. NASA is a premiere scientific and technical institution. World-class research and innovation demands instantaneous access to a wide-array of information sources. Electronic access and delivery is the de facto standard for information providers like the STI Program. Users are not satisfied with the document delivery paradigm. If they cannot retrieve a document from their desktop, they are less likely to pursue obtaining the document through other means.

The high cost of information delivery and increased user demand for it puts information providers in a quandary. To stay relevant and useful to their customers, information providers must explore new ways to capture and disseminate information without straining already limited budgets. For the STI Program, the high cost of maintaining a software and database intensive environment to support information ingestion and dissemination is a driver in the search for less labor intensive and costly technologies to provide access to scientific and technical information. The search for innovative technologies for information delivery has lead to OAI.

Although the OAI movement is relatively new, the momentum for its use is building. Universities and other federal agencies are funding OAI project to test its functionality. OAI takes an uncomplicated approach to interoperability by using Dublin Core as its standard metadata format. Dublin Core is the lowest common denominator for metadata. It lacks the depth of metadata formats like MARC, however, it is not intended to replace native library metadata formats. Its primary purpose is to provide enough information about an object to be useful in a search. For those organizations interested in providing richer metadata formats, OAI supports using additional metadata formats like MARC.

Simplicity is sometimes the best solution to a complicated problem. OAI can be used to build communities of interest that can define their own data exchange practices. OAI offers the flexibility for simple or complex exchange of data. In the case of the TRI Project, participants plan to allow access to full-text PDF documents. The metadata serves as a gateway to the richer data source, the actual technical report. It will depend upon the OAI communities to determine the scope of information exchange.

The STI Program is committed to finding creative and innovative ways to broaden its information delivery services. By participating in the TRI Project experiment and deploying OAI in the TRS environment, the STI Program will be able to evaluate if OAI is a viable solution to some of the challenges of data collection and dissemination. All signs indicate that OAI has great potential for the STI Program efforts.

## REFERENCES


Dublin Core Metadata Initiative (DCMI). "Dublin Core Metadata Element Set, Version 1.1: Reference Description," (2 July 1999); available from http://dublincore.org/documents/dces/; accessed 3 March 2002.

Lynch, Clifford A. "The Z39.50 Information Retrieval Standard - Part I: A Strategic View of Its Past, Present and Future," *D-Lib Magazine* 3, no. 4 (1997); available from http://www.dlib.org/dlib/april97/04lynch.html; accessed 3 March 2002.

Nelson, M.L. et al. "The NASA technical report server," *Internet Research: Electronic Network Applications and Policy* 5, no. 2 (1995); available at http://techreports.larc.nasa.gov/ltrs/papers/NASA-95-ir-p25/NASA-95-ir-p25.html; access 3 March 2002.

Open Archives Initiative, "The Open Archives Initiative Protocol for Metadata Harvesting," Protocol Version 1.0 (21 January 2001), Document Version (24 June 2001); available from http://www.openarchives.org/OAI/openarchivesprotocol.htm; accessed 2 March 2002.

Open Archives Initiative, "The Open Archives Initiative Protocol for Metadata Harvesting," Protocol Version 1.1 (2 July 2001), Document Version (20 June 2001); available from http://www.openarchives.org/OAI_protocol/openarchivesprotocol.html; accessed 2 March 2002.

Open Archives Initiative (OAI). "The Santa Fe Convention: Core Document"; available from http://www.openarchives.org/sfc/sfc.htm; accessed 2 March 2002.

U.S. Congress." National Aeronautics and Space Act of 1958" (PL 85-568, 29 July 1958), 72 Stat., 426, Sec. 203; available from http://www.hq.nasa.gov/office/pao/History/spaceact.html; accessed 7 March 2002.

Van de Sompel, H. and Lagoze, C. "The Santa Fe Convention of the Open Archives initiative," *D-Lib Magazine* 6, no. 2 (2000); available from http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html; accessed 2 March 2002.

Van de Sompel et al."The UPS prototype: an experimental end-user service across e-print archives," *D-Lib Magazine* 6, no. 2 (2000); available from http://www.dlib.org/dlib/february00/vandesompel-ups/02vandesompel-ups.html; accessed 2 March 2002.

Van de Sompel, H. and Lagoze, L."The Open Archives Initiative: Building a low-barrier interoperability framework," *JCDL '01* (25-28 June 2001), Roanoke, VA; available from http://www.cs.cornell.edu/lagoze/papers/oai-final.pdf; access 3 March 2002.

**U.S. Department of Education**
*Office of Educational Research and Improvement (OERI)*
*National Library of Education (NLE)*
*Educational Resources Information Center (ERIC)*

# NOTICE

## Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").